TP : Analyse Statistique des Coûts Médicaux (ANOVA & Chi-2)

Niveau : Licence 3 - Analyse de Données
Durée estimée : 2h00
Outils : Python (Jupyter Notebook)

1. Contexte et Objectifs

Vous travaillez en tant que Data Analyst pour une compagnie d'assurance santé. La direction souhaite ajuster ses primes d'assurance. Pour cela, elle a besoin de comprendre comment les caractéristiques des assurés (âge, région, tabagisme...) influencent leurs frais médicaux (charges) et s'il existe des corrélations entre les profils démographiques.

Le Dataset :
Le fichier contient 1338 lignes et 7 colonnes :

- age : Âge de l'assuré.
- sex : Sexe (female, male).
- bmi : Indice de masse corporelle (numérique).
- children : Nombre d'enfants couverts.
- smoker : Fumeur (yes, no).
- region : Zone de résidence aux USA (northeast, northwest, southeast, southwest).
- charges : Frais médicaux facturés par l'assurance (Variable cible).

Lien de téléchargement des données (Source brute) :
https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insurance.csv

Consigne :

1. Importez le fichier CSV dans votre environnement.
2. Vérifiez la structure des données (types de variables) et la présence éventuelle de valeurs manquantes (NA).
3. Affichez les statistiques descriptives de base (describe() sous Python).

2. Étude de Cas A : Le BMI dépend-il de la région ? (ANOVA)

On cherche à savoir si l'Indice de Masse Corporelle (bmi) moyen est significativement différent selon les 4 régions (region).

Question 1 : Analyse Visuelle (Validation intuitive)

Tracez une **Boîte à Moustaches (Boxplot)** croisant la region (axe X) et le bmi (axe Y).

Observation : Regardez la taille des boîtes (la hauteur). Semblent-elles avoir grossièrement la même taille ?

Note : Si les boîtes ont des tailles comparables, on peut supposer que la variabilité (variance) est stable entre les groupes, ce qui nous autorise à faire l'ANOVA sans calculs complexes.

Question 2 : Statistiques Descriptives

Calculez la **moyenne** et l'**écart-type** du BMI pour chaque région.

Vérification : Les écarts-types sont-ils proches les uns des autres ? (Par exemple, s'ils sont tous autour de 6.0, c'est parfait).

Question 3 : Test ANOVA

Posez vos hypothèses :H_0 : Le BMI moyen est identique partout.

H_1: Au moins une région a un BMI moyen différent.

Lancez la commande d'ANOVA (One-Way).

Analysez la **p-value**. Si elle est inférieure à 0.05, rejetez H_0.

Question 4 : Qui est différent ? (Post-hoc)

Conclusion business : Quelle est la région la plus "à risque" en termes de surpoids ?

3. Étude de Cas B : Le tabagisme dépend-il de la région ? (Chi-2)

On cherche à savoir s'il existe une dépendance entre la zone d'habitation (region) et le fait d'être fumeur (smoker).

Question 1 : Tableau Croisé

Générez le tableau de contingence (Tableau croisé dynamique) comptant le nombre de fumeurs et non-fumeurs par région.

Question 2 : Test du Chi-2 (Chi-Carré)

Posez vos hypothèses :H_0 : Il y a indépendance (la proportion de fumeurs ne dépend pas de la région).

H_1 : Il y a dépendance.

Lancez le test du Chi-2. Regardez la **p-value**. Est-elle inférieure à 0.05 ?

Question 3 : Interprétation

Que concluez-vous sur le lien entre la région et le tabagisme ? Est-il statistiquement significatif sur cet échantillon ?

4. Synthèse Managériale

Rédigez 3 phrases pour votre directeur :

Le lieu d'habitation a-t-il un impact sur l'obésité (BMI) ?

Le lieu d'habitation a-t-il un impact sur le tabagisme ?

Recommandez-vous de moduler les prix de l'assurance selon la région ?